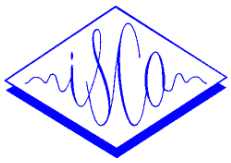
ISCA Archive
http://www.isca-speech.org/archive



3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

PERCEPTION FOR VCV SPEECH UTTERED SIMULTANEOUSLY OR SEQUENTIALLY BY TWO TALKERS#

Kazuhiko Kakehi and Kazumi Kato***

*Graduate School of Human Informatics, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-01 Japan
**NTT Basic Research Laboratories
3-1 Wakamiya, Morinosato, Atsugi, Kanagawa, 243-01 Japan

ABSTRACT

We investigated robust perception of speech using different talkers VCVs uttered simultaneously or sequentially. We conducted two experiments: one to investigate the perception of the number of talkers and the other to investigate the perception of intervocalic consonants. The experimental results are as follows : (1) Even when two talkers are perceived in a VCV stimulus, the extracted phoneme features are perceptually integrated regardless of the talker source information. (2) Phoneme features in a pre-closure part in double talk contribute to intervocalic stop consonant perception regardless of talker source information. This means that the talker source information is used to determine speech portion where phoneme features are extracted, and that these features are integrated to perceive a phoneme regardless of the talker source information.

I. INTRODUCTION

Human beings have the ability to perceive and recognize speech robustly in various environments. It is well known that we can communicate with each other in very noisy environments in which many people are talking back and forth. This is known as the cocktail party effect. Studies show that the cocktail party effect is a total effect from lower to higher processing of speech perception [1]. In this paper we investigate the role of talker source information in the lower stage of speech processing to clarify its effect on phoneme/syllable perception in mixed speaker conditions.

We studied the perception of VCV speech, uttered sequentially by two talkers, to show that there are at least two stages in the process of perceiving phonemes. They are the phoneme cues extraction and integration stages. We further investigated the process of perception using stimuli of VCV speech uttered by two talkers simultaneously or sequentially.

II. EXPERIMENTS

We conducted two experiments: Experiment 1 is on the perception of a stop consonant in VCV stimuli uttered by two talkers simultaneously or sequentially, and Experiment 2 is on the perception of the number of talkers using the same kind of stimuli as in Experiment 1.

Stimuli

Stimuli used in this experiment were VCVs. The Vs are the five Japanese vowels (/a,e,i,o,u/) and the Cs are voiceless stop consonants (/p,t,k/).

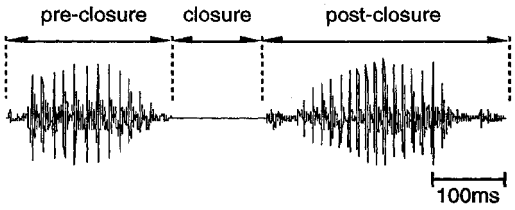


Fig. 1. A waveform example of VCV (/apa/) utterance.

#) This study was conducted in NTT Basic Reserach Laboratories.

Figure 1 shows an example of the waveform for a VCV utterance in which a clear silent interval closure part is observed separating the VCV into a pre-closure part and a post-closure part.

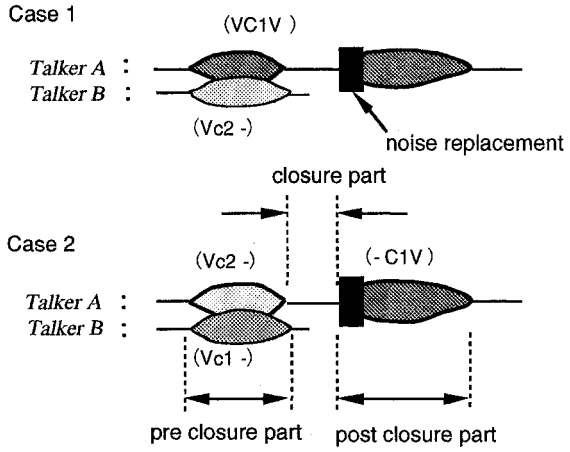


Fig.2 Stimuli used in the Experiment 1 & 2.

Figure 2 shows the basic structures of stimuli illustrated with wave envelopes of the VCV utterances. Stimuli consist of mixtures of two talker's utterances in which the pre-closure part (Vc-) is concatenated with the post-closure part of a VCV uttered by one of the two talkers. In case 1, consonant information in the pre-closure part and the post-closure part is consistent within the VC1V uttered by talker A, but inconsistent between the utterance of talker A and that of talker B. In case 2, consonant information is inconsistent between the closure part of talker B (Vc1-) and the post-closure part of talker A (-C1V). In these cases, the cross-spliced part is taken from the VCV utterances which have the same vowel combinations from the five Japanese vowels.

The differences in utterance levels between talkers A and B in the pre-closure part were varied from $-\infty$ to ∞ in terms of signal-to-noise rates. Two level conditions were used, that is, the level of talker A reduced while the level of talker B is kept at the originally uttered level uttered and vice versa.

In total, there were four stimuli conditions made by combining the information consistency conditions with the level conditions.

As the phonemic cues in the burst and formant transition part of stop consonants are very strong, the initial portion of the post-closure part was replaced by noise up to 70ms to clearly examine the perceptual integration of phonemic cues distributed in the time domain.

Experiment 1:

Four subjects trained for sound listening tests listened diotically to the VCV stimuli with a headphone set and wrote down the syllables they heard. Two female talkers were used to make the stimuli as described above. The level of replacement noise was 5 dB larger than the VCV utterance level.

The listening level was set at 70dB SPL based on the subjects' preference. The intermittent stimulus interval was approximately 4 seconds and included a 1 kHz tone signal 50ms long.

Results of Experiment 1

Figure 3 shows the experimental results in terms of consonant perception scores. For the C2 perception curves , the abscissa means the signal-to-noise ratio, eg. , a talker's utterance that includes C1 information is taken for noise while a talker's utterance that includes C2 information is taken for signal. For the C1 perception curves, which is the inverse of the signal-to-noise ratio, the squares show the results for the stimuli of case 1 and the circles show the results for the stimuli of case 2.

In each condition, the sum of the perception scores of C1 and C2 is nearly 100%. This means that the subject did not perceive a consonant of which the information was not included in the VCV stimuli. The results show very clearly that the perception scores of C1 or C2 are determined by the simple parameter of signal-to-noise ratio regardless consistency of talker source information between a pre-closure part and a post-closure part.

When the utterance levels of two talkers are equal or 0dB in abscissa, the phonemic information included in each talker's utterance was used equally. In this case the consonant perception score for C1 and C2 is 50% each. When the level difference of the two talkers' utterance is greater than 15dB, the utterance of the smaller level has no effect on perception.

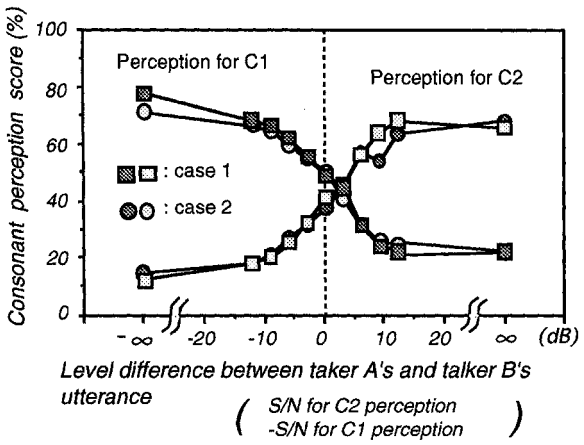


Fig.3 Consonant perception score in Experiment 1.

Experiment 2

Subjects were instructed to state the number of talkers and syllables perceived for the same kind of stimuli used in the experiment 1. The stimuli were made from three talkers' utterance (two female, F1 and F2, and one male, M1). The preceding and succeeding vowels to the intervocalic stop consonants were fixed at /a/ and /e/respectively in Experiment 2. The talker combinations were set as follows: (1) Adding talker B's pre-closure part of VCV utterance to talker A's VCV. (No phoneme cues were inconsistent in talker A, but there was inconsistency between talker B's pre-closure part and talker A's post-closure part. (2) The same condition as (1), except talker A's VCV utterance was cross-spliced (phoneme cues were inconsistent for talker A, but consistent between talker B and talker A), and (3) Adding talker A's pre-closure part of the VCV utterance to the same talker A's VCV utterance. Both stimuli with and without noise replacement were used. Two groups of subjects were used.

Two subjects in Group 1(G1) were trained for syllable listening tests and are familiar with the talker's voice used. Four subjects in Group 2(G2) were untrained subjects and have no experience to hear the talker's voice before the experiment. Subjects heard the stimuli diotically at their preferred listening level. They wrote down the number of talkers their perceived first and then the syllables.

Results of Experiment 2

Subject is response for the number of perceived talkers was almost always one or two. Responses of three are rare. There is no noise replacement effect on the number of talkers perceived. Figure 4 shows the perception rate of two talkers for stimuli conditions (1) and (2), respectively.

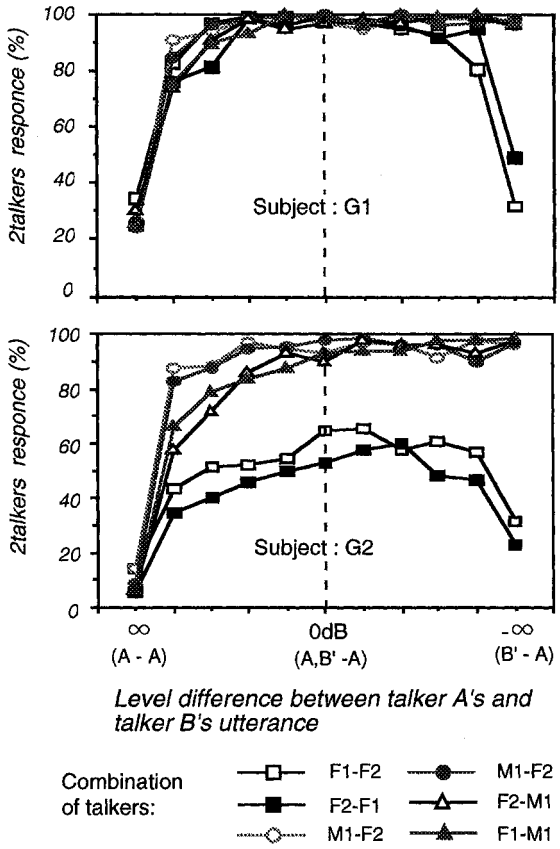


Fig.4 Number of talkers percieved in Experiment 2.

The abscissa indicates the level difference of talker A and B. (the left side (A-A) means no B's utterance, the middle (A,B-A) means that the level of talker A's utterance is equal to talker B's, and the right side (B'-A) means no A's utterance in the pre-closure part). The experimental results for the stimuli conditions (1) and (2) are nearly the same as with Group 2; there is no clear difference except for the responses to the stimuli for the female utterance combinations. Untrained subjects have a tendency to hear two female talkers' utterance as one.

For stimuli condition (3), which includes only one talker's utterance, a large number of

subjects responded that there were 2 talkers, especially for the female utterance stimuli. In every case, consistency/inconsistency of the phonemic information has no effect on the number of talkers perceived.

III. DISCUSSION

We have investigated the perceptual integration of phonemic cues distributed in time domain, using a cross-spliced VCV stimuli which includes phonemic and talker source information conflict in reference [2]. We showed that there are at least two stages in the processing of phoneme/syllable perception from the point of view for talker source information used. The first stage is the identification of the sound source and the succeeding extraction of phoneme cues from the same sound source where the talker source information is used. The second stage is the perceptual integration of the once extracted phonemic cues regardless of the talker source information. The stimuli used in reference [2] were speech uttered by one talker or two talkers cross-spliced.

In a usual speech environment, we do not have the experience often to hear speech uttered by two talkers sequentially just in good timing. So, in this paper, we used VCV speech uttered simultaneously or sequentially by two talkers, which is the usually case.

Experiment 1 shows very clear results that the perception rate of the consonant is only determined by the difference in the level of two talker's utterances. It indicates that the chance of extraction of phonemic cues is only determined by the difference in the level of two talker's simultaneous utterance and that the once extracted phonemic cues are perceptually integrated regardless of talker source information.

Question might be raised concerning the fact that the talkers used in Experiment 1 were both female. If the talker source information is different, such as in the combination of a female and male voice, would we get the same results? We conducted an additional experiment using male and female talkers, and the results are almost the same as in Experiment 1.

Cutting conducted a dichotical experiment in which the listeners got one formant of a syllable to one ear and second formant to the other ear. When the two formants had different fundamental frequencies, listeners heard two sound sources but were still able to perceive the syllable[3]. In experiment 1, the same kind of phenomenon was observed in diotic listening.

Darwin showed that the fundamental frequencies of formants are used for perceptual integration (grouping) to perceive syllables in critical cases (for example, some formants were in common between two syllables at the same time)[4].

The results of Experiment 1 do not seem to match with Darwin's results. This is attributed to the difference of grouping phonemic cues, that is, perceptual integration is sequential time domain in Experiment 1, which is simultaneous in Darwin's experiment.

Experiment 2 shows the number of talkers perceived by the subjects during phoneme/syllable perceptual processing. The talker source information is used for the perception of the number of talkers. These results indicate that even if we are clearly aware of the number of talkers, the phonemic cues are perceptually integrated regardless of talker source information.

V. CONCLUSION

That the results strongly support the hypothesis that talker source information is used to determine speech portion where phoneme cues are extracted. These features are integrated to perceive a phoneme regardless of the talker source information.

REFERENCES

[1] E. C. Cherry and R. Wiley, "Speech communication in very noisy environments," *Nature 214, 1164 (1976).*

[2] K. Kato and K. Kakehi, , "A role for talker source information in the perception of intervocalic stop consonants," *JASJ 50, 83-90 (1994)*

[3] J. E. Cutting, "Auditory and linguistic process in speech perception: Inferences from six fusions in dichotic listening," *Psychol. Rev. 83 (2), 114-140 (1976)*

[4] C. J. Darwin, "Perceptual grouping of speech components differing in fundamental frequency and onset time," *Q. J. Exp. Psychol. 33A, 185-207 (1989)*